AMS pre-conference workshop
23rd Jan. 2010

Verification of ensemble streamflow forecasts using the Ensemble Verification System (EVS)

James Brown, Julie Demargne, OHD

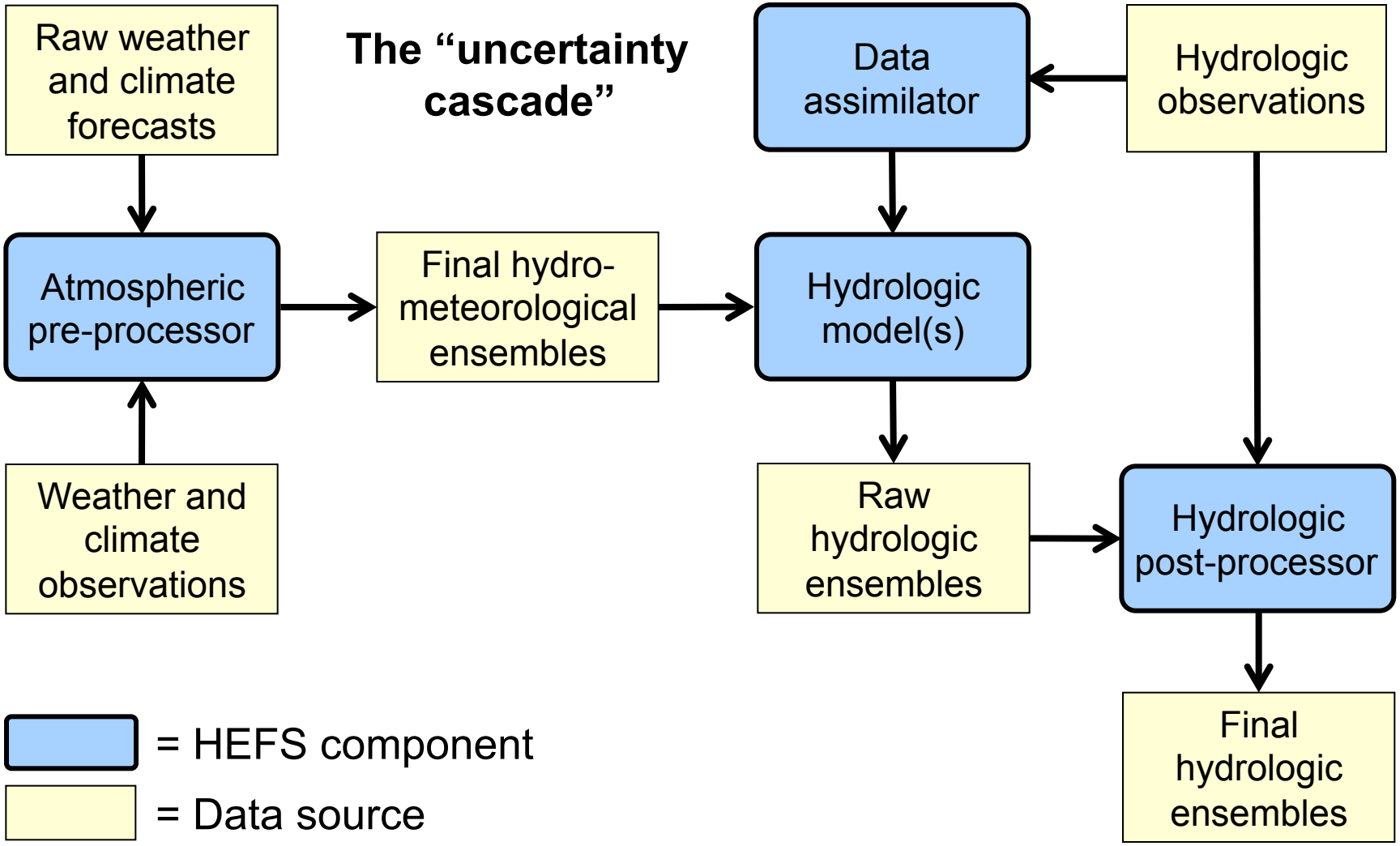
james.d.brown@noaa.gov

Overview

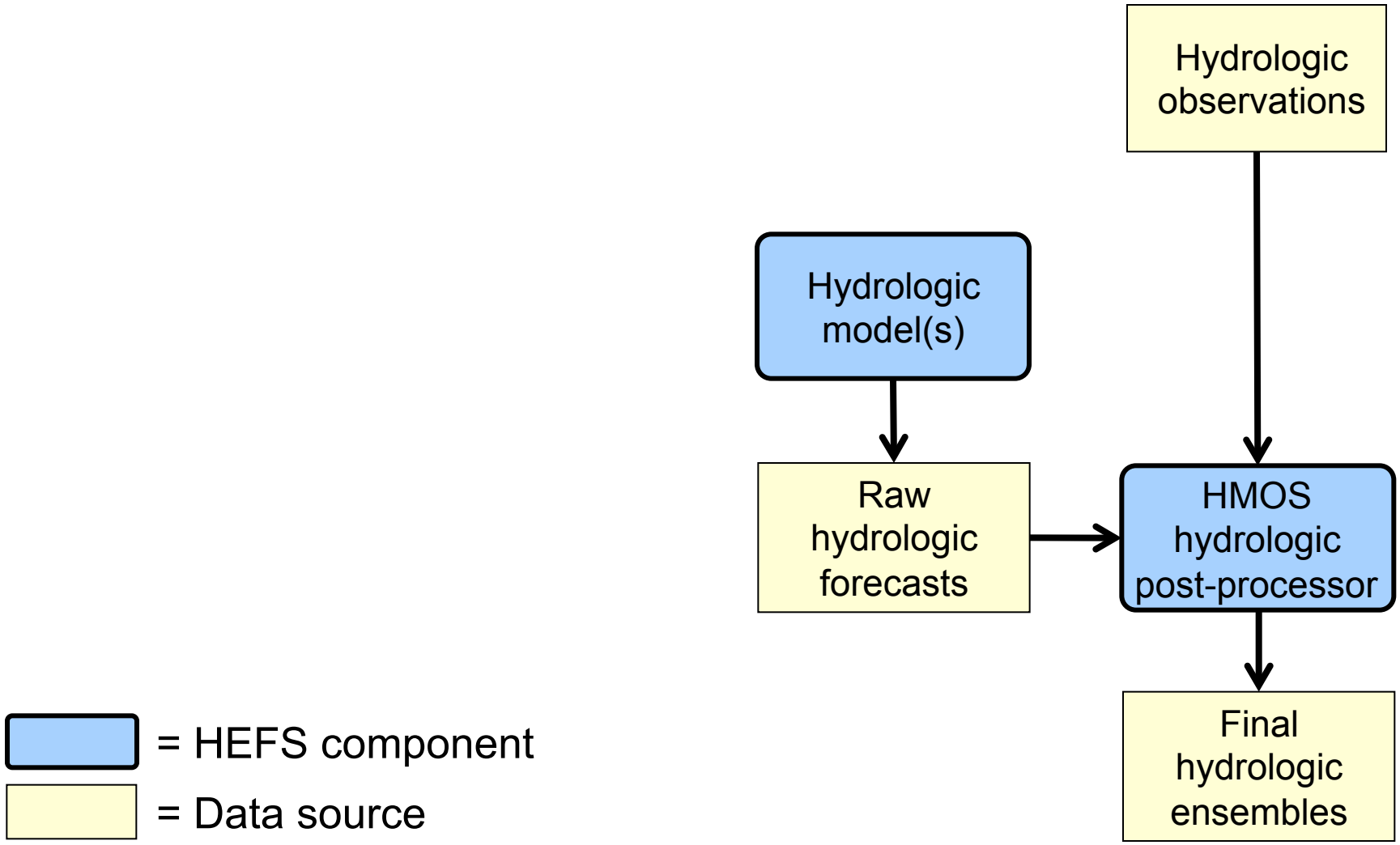
1. **Brief review of the NWS HEFS**
   - Two approaches to generating ensembles
   - "Bottom-up" (ESP) vs. "top down" (HMOS)
2. **Verification of streamflow ensembles**
   - Techniques and metrics
   - Ensemble Verification System (EVS)
3. **Example: ESP-GFS from CNRFC**

1.Brief review of the NWS HEFS

Bottom-up ("ESP")



Top down (HMOS)



Pros and cons of "ESP"

Pros

- Knowledge of uncertainty sources
- Can lead to targeted improvements
- Dynamical propagation of uncertainty

Cons

- Complex and time-consuming
- Always residual bias (need post-processing)
- Manual intervention is difficult (MODs)

Pros and cons of HMOS

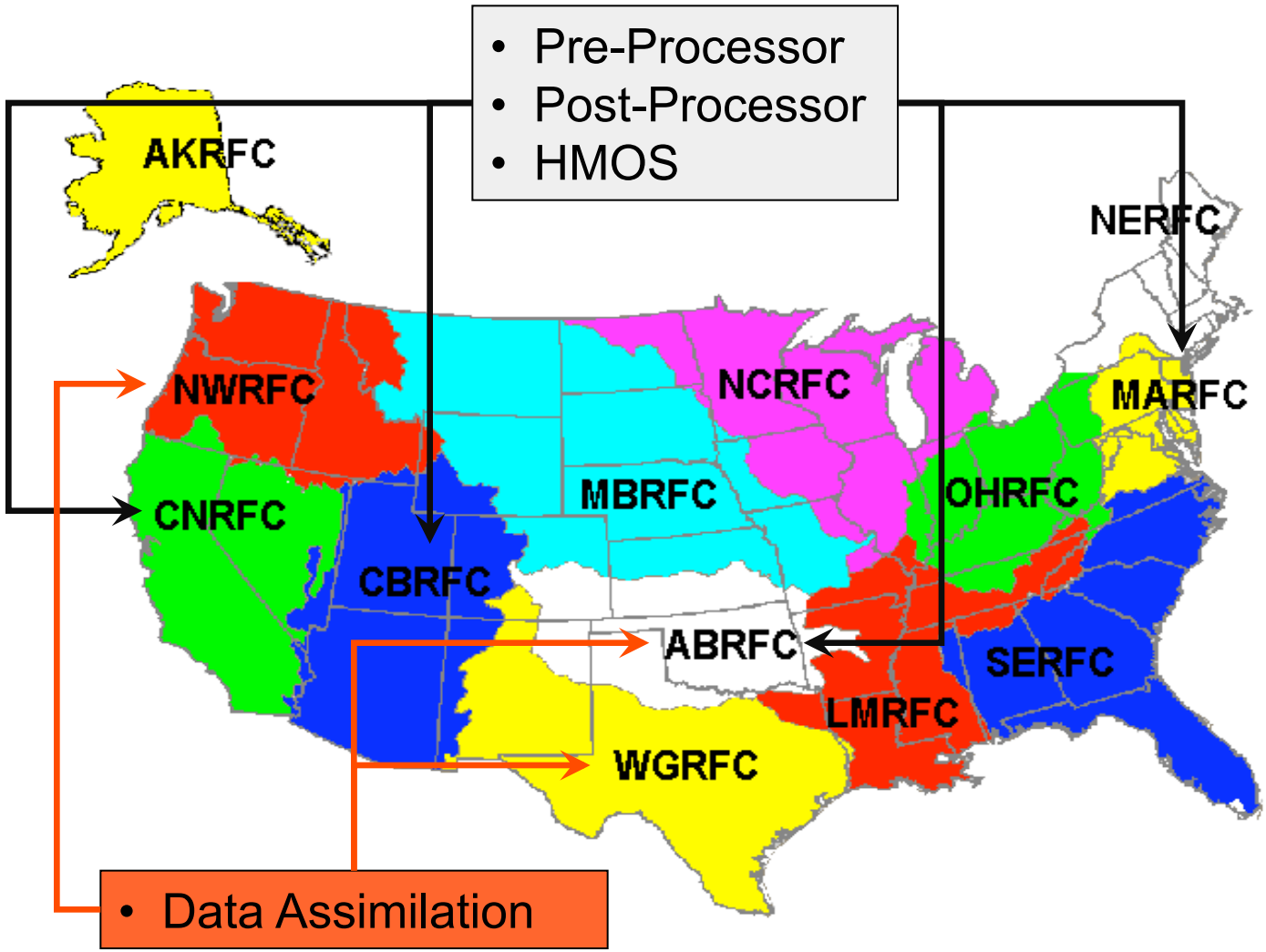
Pros

- Simple statistical technique
- Produces reliable ensemble forecasts
- Uses single-valued (e.g. MOD'ed) forecasts

Cons

- Requires statistical assumptions
- Benefits are often short-lived (correlation)
- Lumped treatment (no source identification)

Status of X(H)EFS testing



2. Verification of streamflow ensembles

A “good” flow forecast is..?

Statistical aspects

- Unbiased (many types of bias….)
- Sharp (doesn’t say “everything” possible)
- Skilful relative to baseline (e.g. climatology)

User aspects (application dependent)

- Sharp
- Warns correctly (bias may not matter)
- Timely and cost effective

Statistical aspects

Distribution-oriented verification

- Q is streamflow, a random variable.
- Consider a discrete event (e.g. flood): $\{Q > q_v\}$.
- Forecast (y) and observe (x) many flood events.

$$y_i = \Pr[Q > q_v], \quad x_i = \{1 \text{ if } Q > q_v, \text{ else } 0\} \quad i = 1,\ldots,n$$

How good are our forecasts for $\{Q>q_v\}$?

- Joint distribution of forecasts and observations
- $f(x,y) = a(x|y) \cdot b(y)$ "calibration-refinement"
- $f(x,y) = c(y|x) \cdot d(x)$ "likelihood-base-rate"

(Some) attributes of quality

Calibration-refinement: $a(x|y) \cdot b(y)$

- **Reliable if (e.g.):** $E[x \mid y = p] = p \quad \forall p$
- **"When $y = 0.2$, should observe 20% of time"**
- **Sharp if:** $y \rightarrow 0$ or 1
- **"Maximize sharpness subject to reliability"**

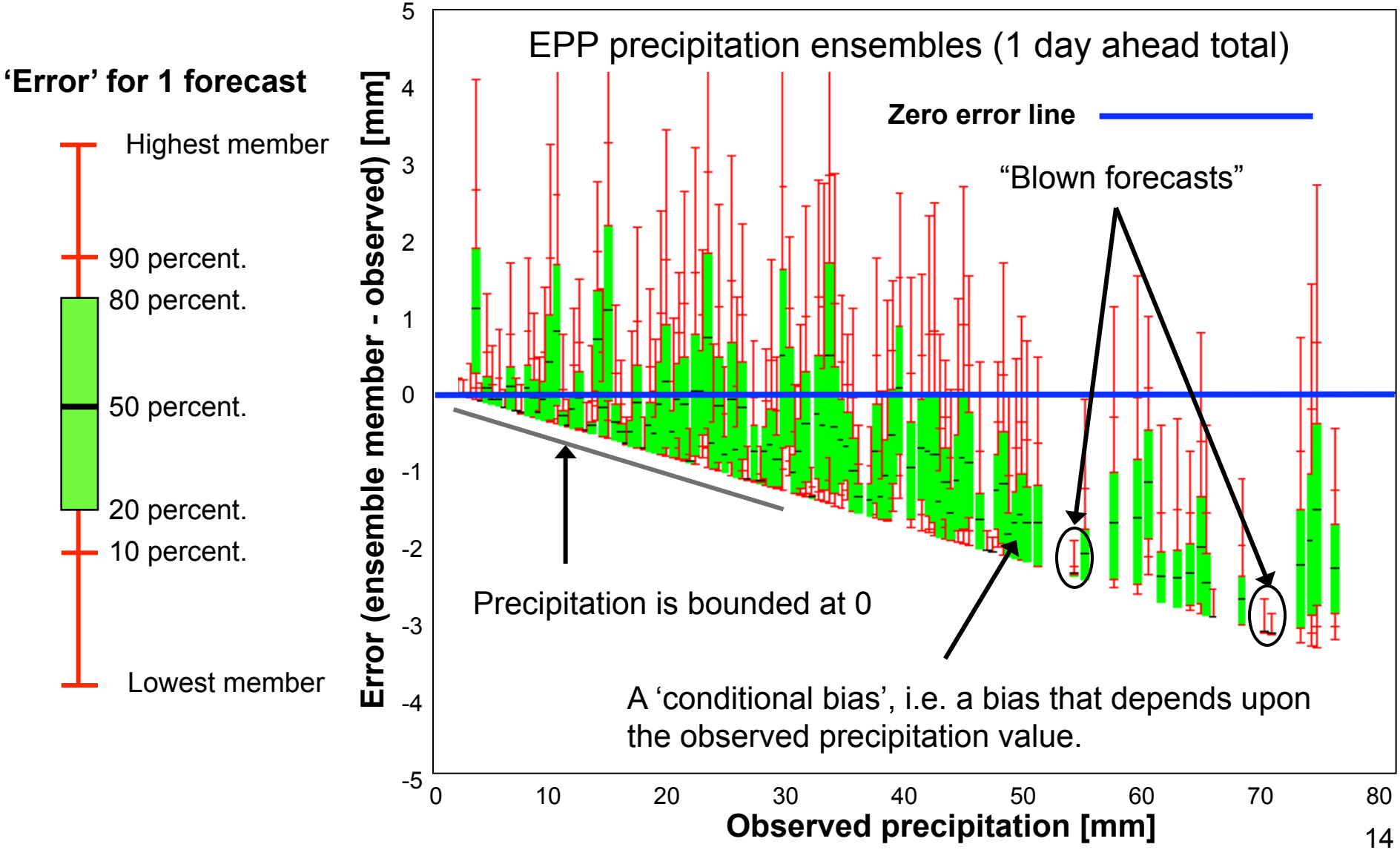
Likelihood-base-rate: $c(y|x) \cdot d(x)$

- **Discriminatory if (e.g.):**
  $E[y \mid x = 1] >> E[y \mid x = 0]$
- **"Forecasts easily separate flood from no flood"**

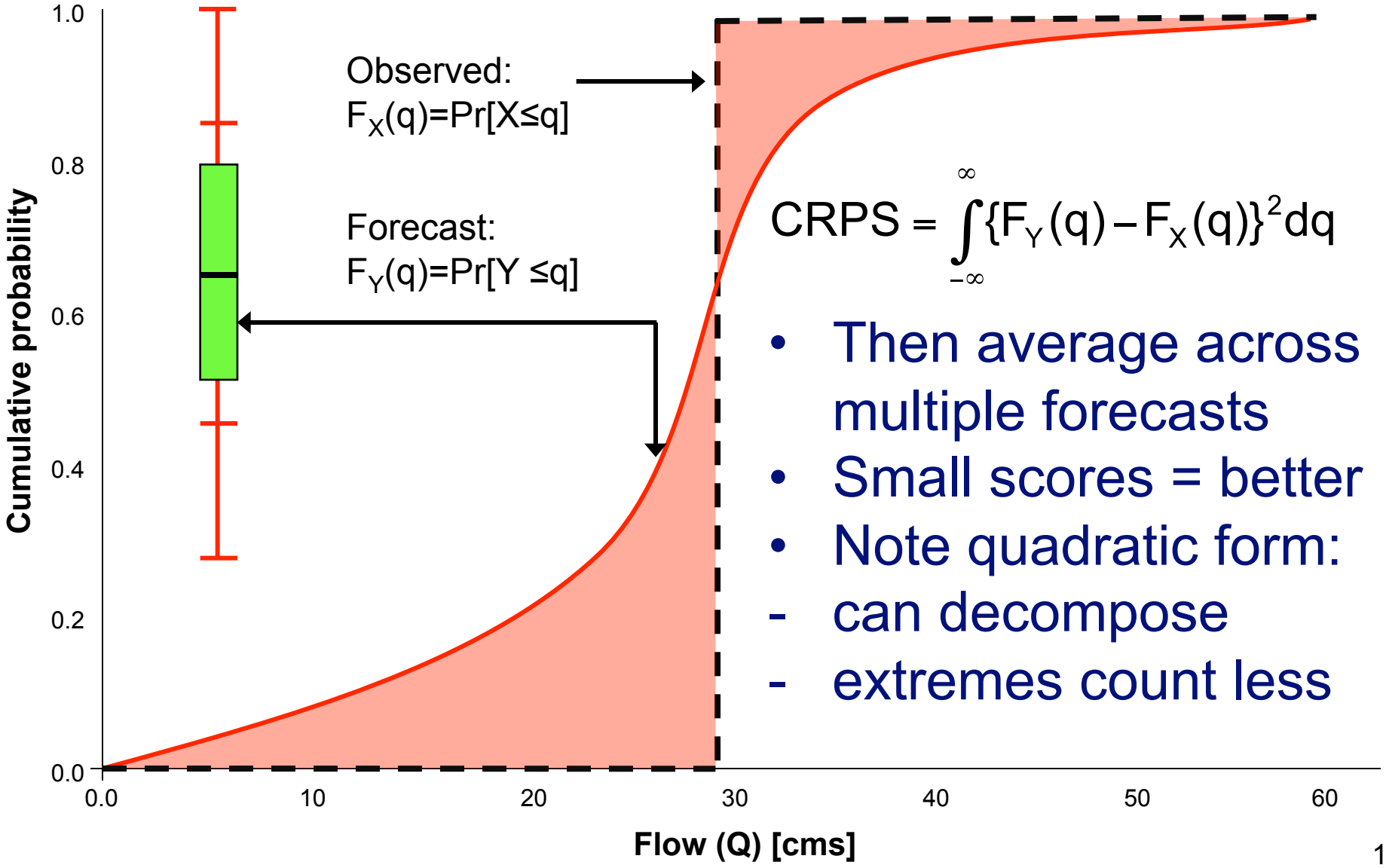
(Some) quality metrics

1. **Exploratory metrics (plots of pairs)**
2. **Lumped metrics or 'scores'**
   - Lumps all quality attributes (i.e. overall error)
   - Often lumped over many discrete events
   - Include skill scores (performance over baseline)
3. **Attribute-specific metrics**
   - Reliability Diagram (reliability and sharpness)
   - ROC curve (event discrimination)

Exploratory metric: box plot

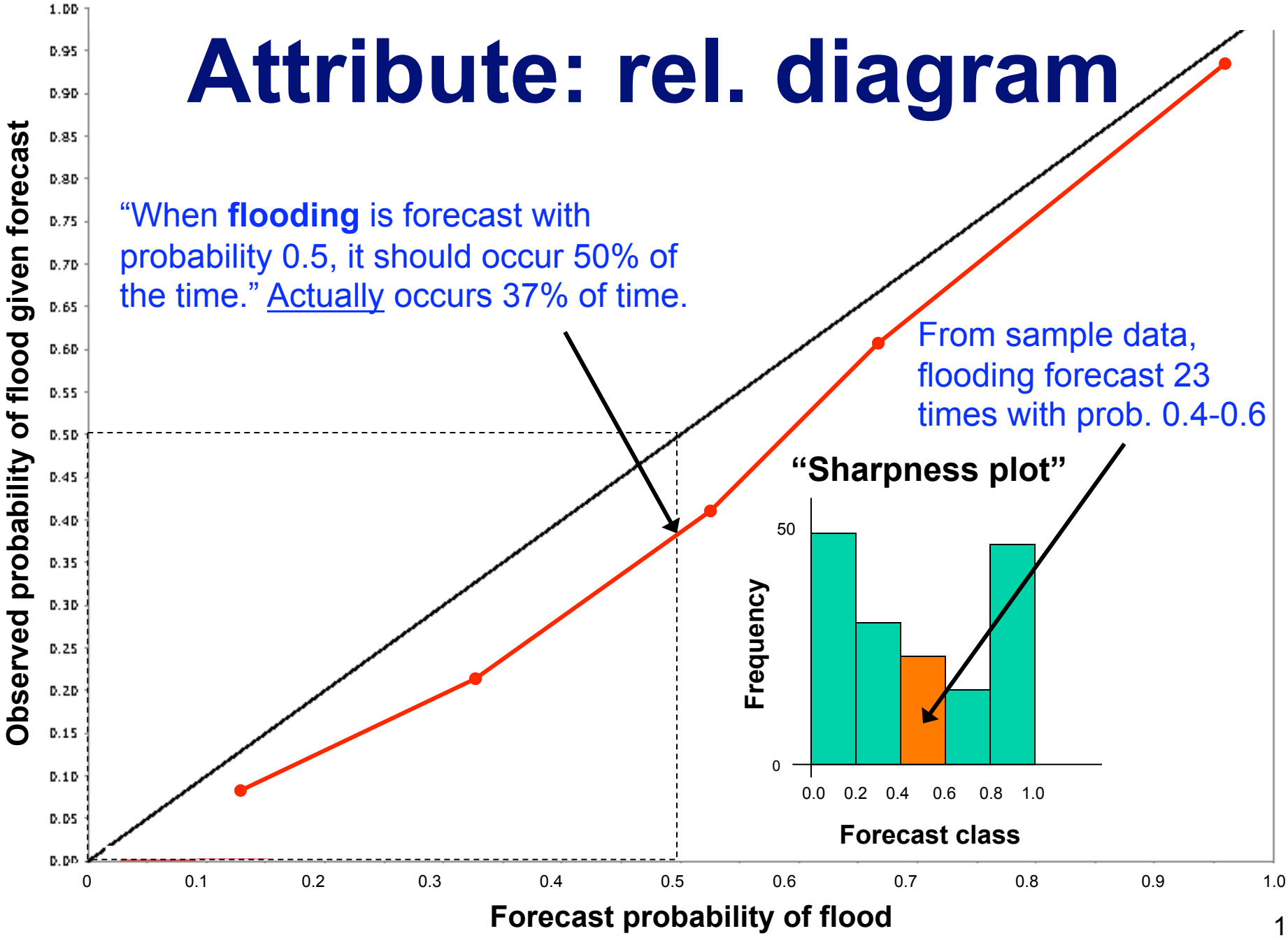


Lumped metric: Mean CRPS



- Then average across multiple forecasts
- Small scores = better
- Note quadratic form:
  - can decompose
  - extremes count less

Attribute: rel. diagram



The Ensemble Verification System (EVS)

The EVS

Java-based tool

- GUI and command line. GUI is structured….

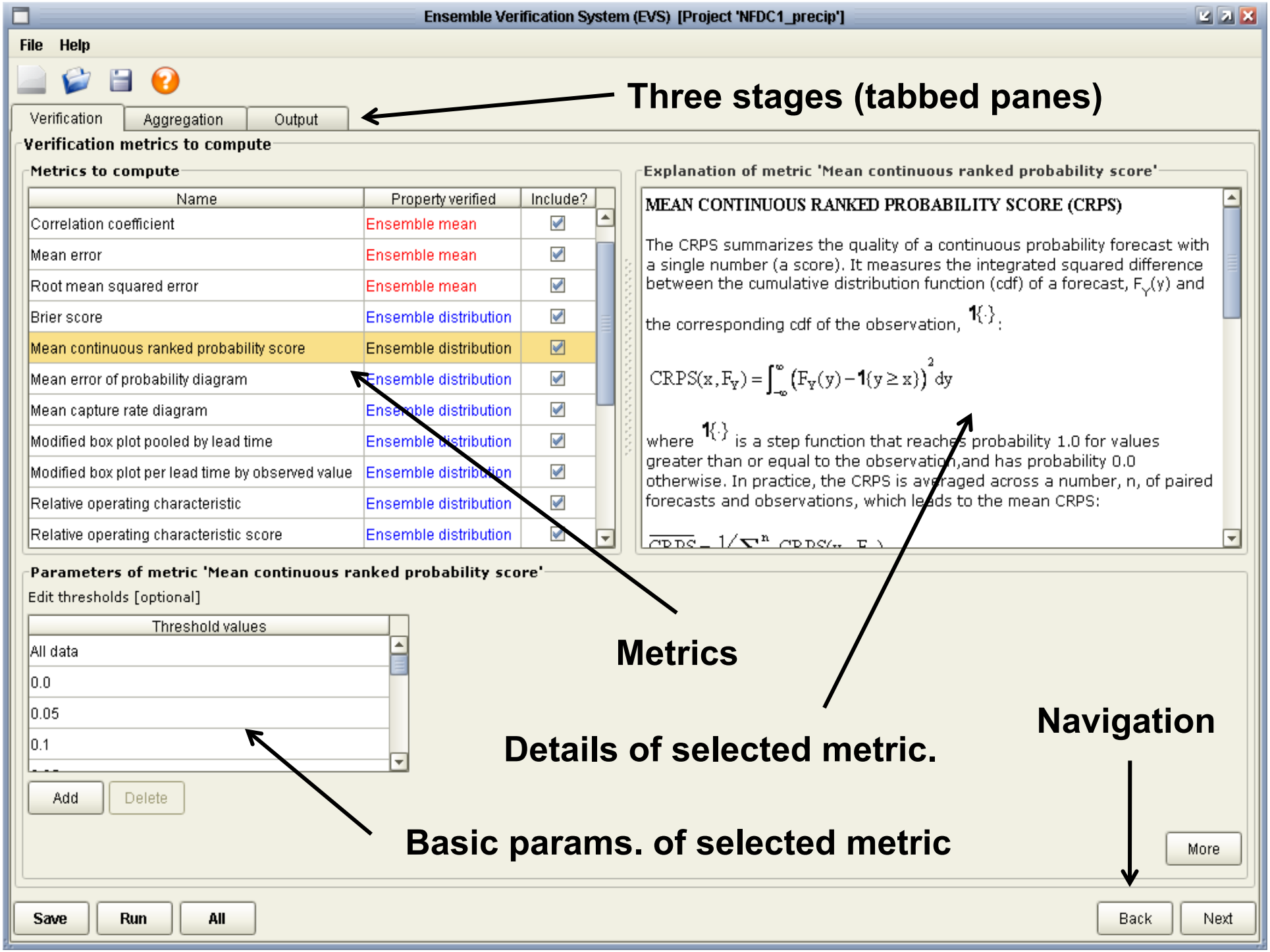
1. Verification (at specific locations)

- Add locations, data sources, metrics etc.

2. Aggregation (across locations)

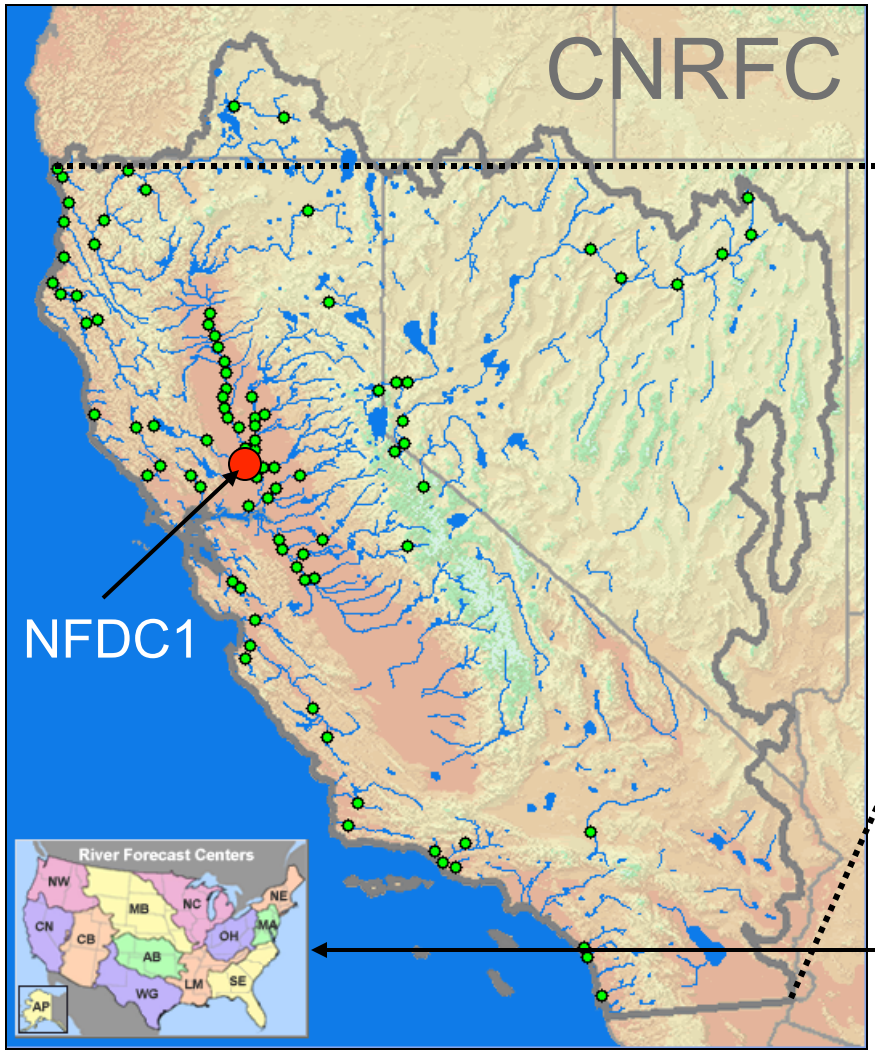
- Compute aggregate performance

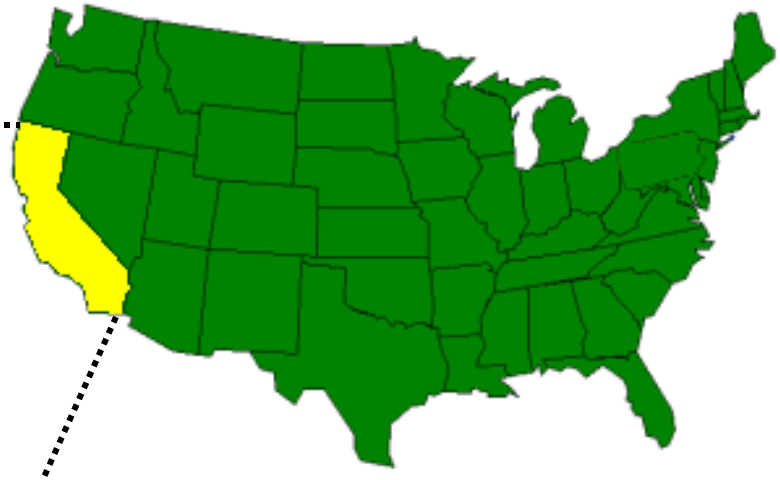
3. Output (graphical and numerical)


Ensemble Verification System (EVS) [Project 'NFDC1_precip']
File Help
Three stages (tabbed panes)
Verification Aggregation Output
Verification metrics to compute
Metrics to compute
Name | Property verified | Include?
Correlation coefficient | Ensemble mean
Mean error | Ensemble mean
Root mean squared error | Ensemble mean
Brier score | Ensemble distribution
Mean continuous ranked probability score | Ensemble distribution
Mean error of probability diagram | Ensemble distribution
Mean capture rate diagram | Ensemble distribution
Modified box plot pooled by lead time | Ensemble distribution
Modified box plot per lead time by observed value | Ensemble distribution
Relative operating characteristic | Ensemble distribution
Relative operating characteristic score | Ensemble distribution
Explanation of metric 'Mean continuous ranked probability score'
MEAN CONTINUOUS RANKED PROBABILITY SCORE (CRPS)
The CRPS summarizes the quality of a continuous probability forecast with a single number (a score). It measures the integrated squared difference between the cumulative distribution function (cdf) of a forecast, F_Y(y) and the corresponding cdf of the observation, 1{.}:
CRPS(x, F_Y) = ∫_{-∞}^{∞} (F_Y(y) − 1(y ≥ x))^2 dy
where 1{.} is a step function that reaches probability 1.0 for values greater than or equal to the observation,and has probability 0.0 otherwise. In practice, the CRPS is averaged across a number, n, of paired forecasts and observations, which leads to the mean CRPS:
Parameters of metric 'Mean continuous ranked probability score'
Edit thresholds [optional]
Threshold values
All data
0.0
0.05
0.1
Add Delete
Metrics
Details of selected metric.
Navigation
Basic params. of selected metric
More
Save Run All
Back Next

3. Example application

N. Fork, American (NFDC1)





NFDC1: dam inflow. Lies on upslope of Sierra Nevadas.

13 NWS River Forecast Centers

Data available (NFDC1)

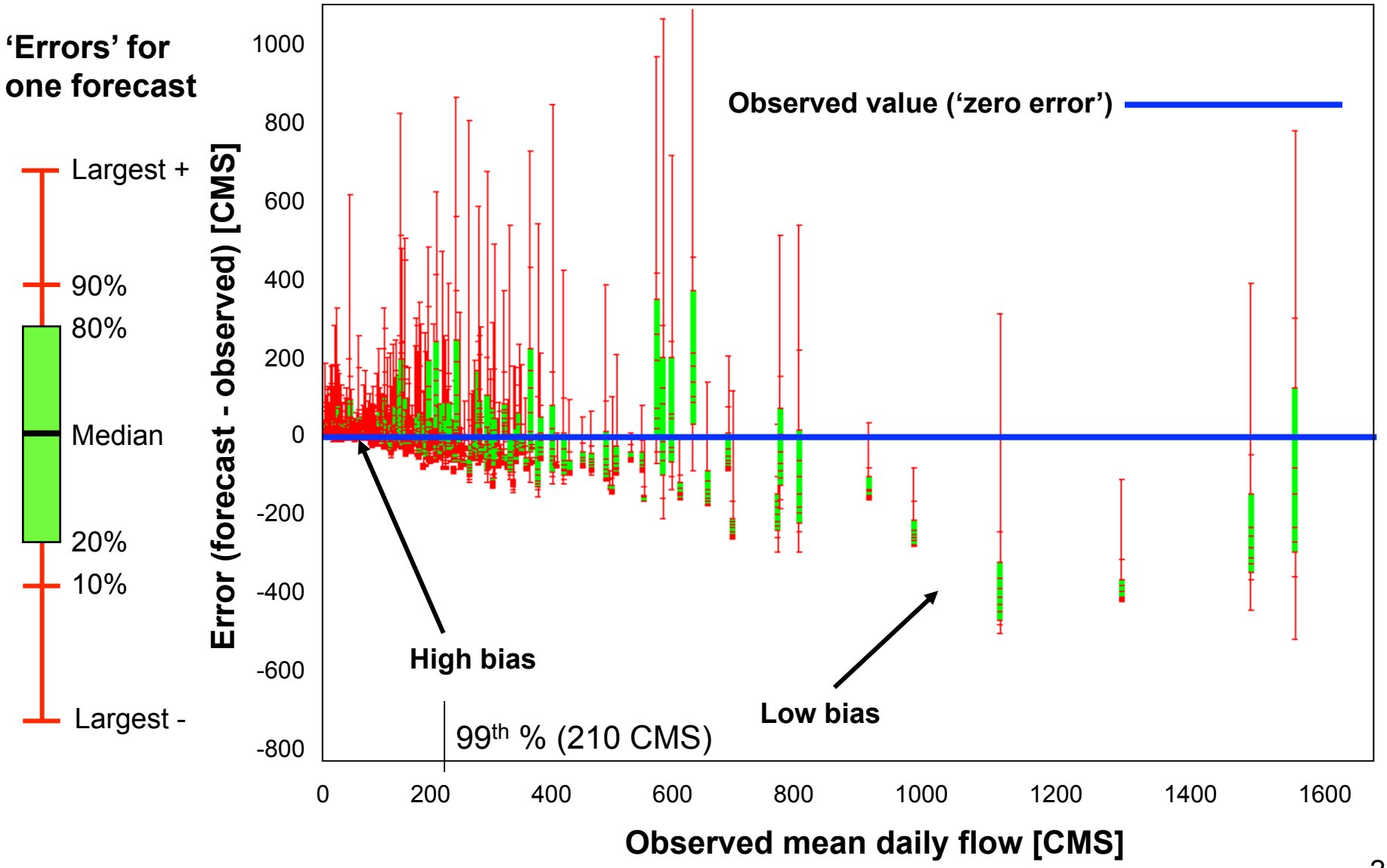
Streamflow ensemble forecasts

- Ensemble Streamflow Prediction system
- NWS RFS (SAC) w/ precip./temp. ensembles
- Hindcasts of mean daily flow 1979-2002
- Forecast lead times 1-14 days ahead
- NWS RFS (SAC) is well-calibrated at NFDC1

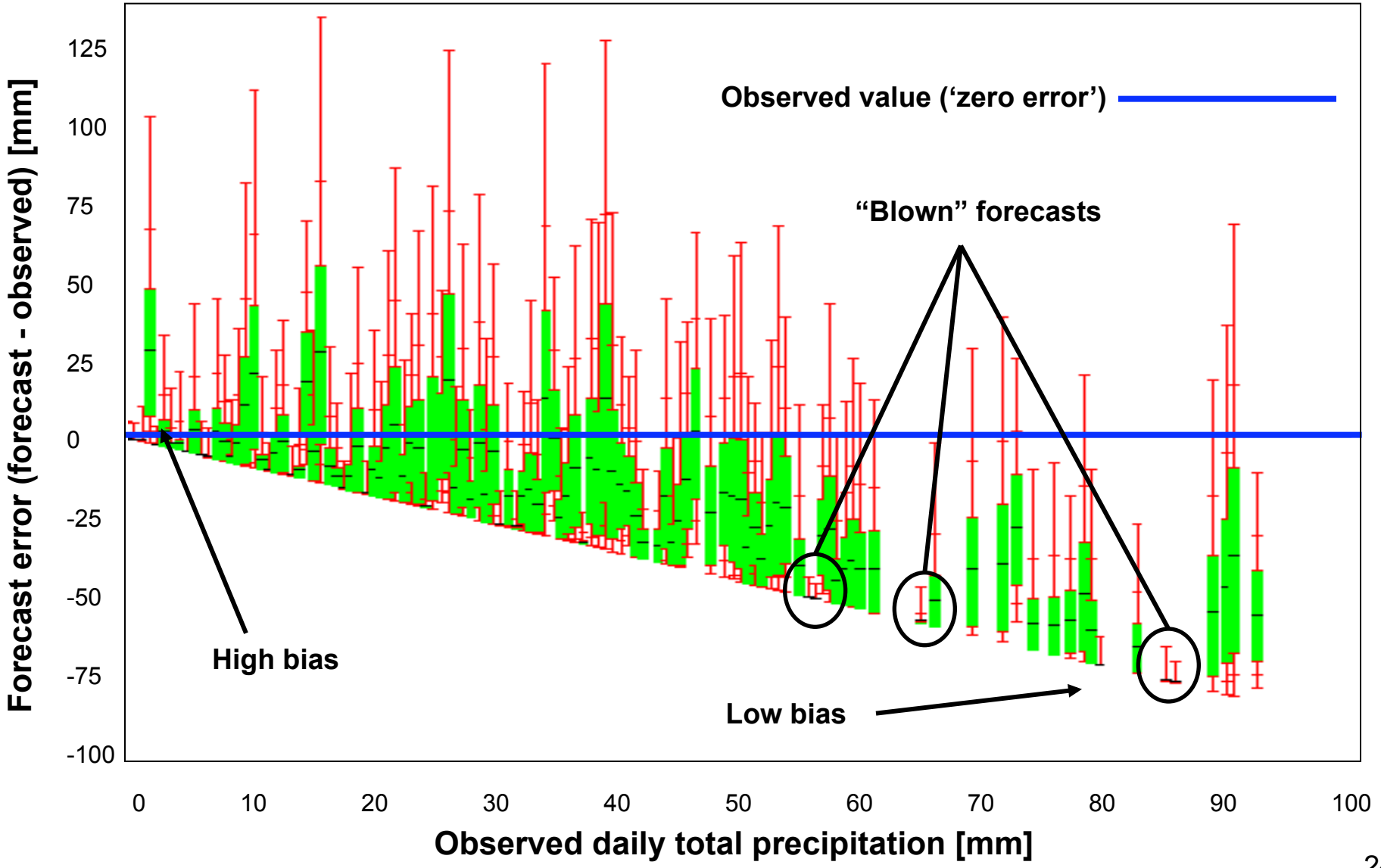
Observed daily flows

- USGS daily observed stage
- Converted to discharge using S-D relation

Box plot of flow errors (day 1)

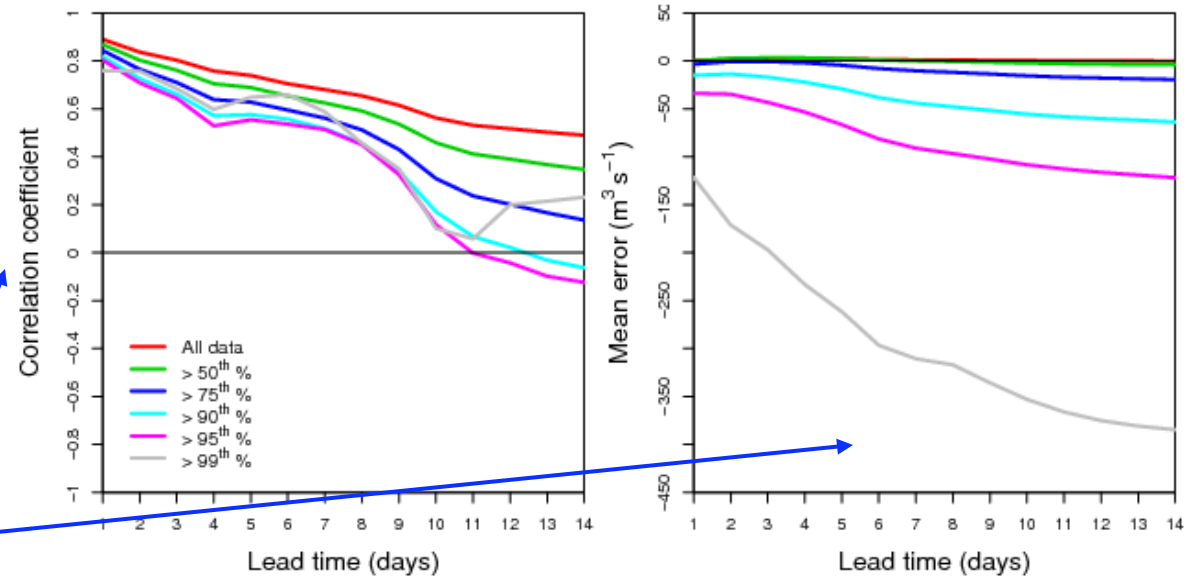


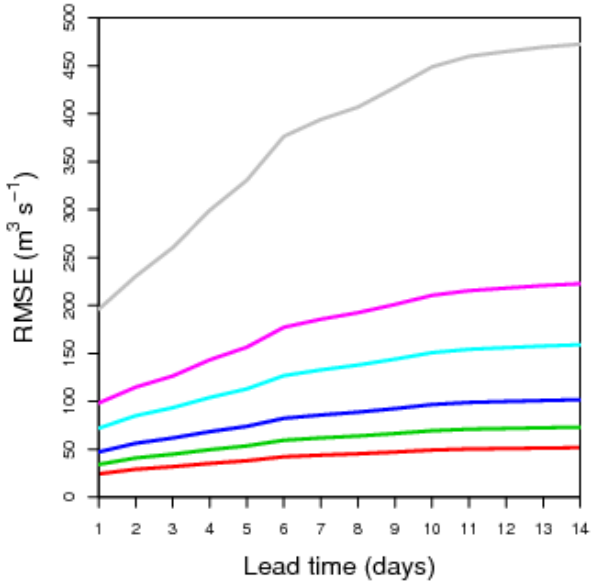
Precipitation (day 1, NFDC1)

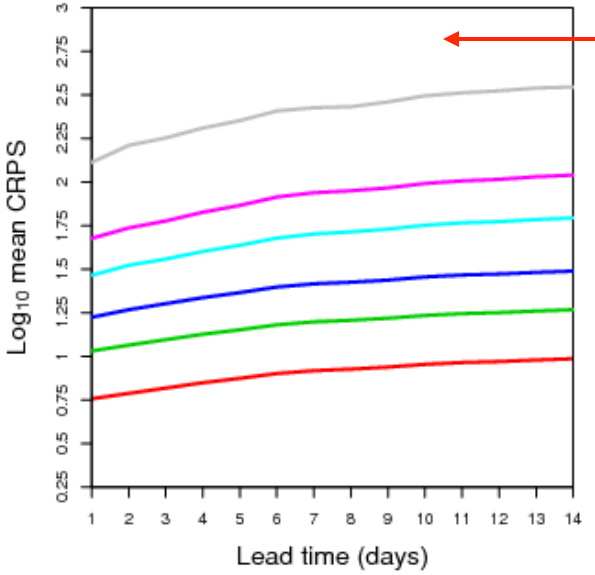


Lumped error statistics

Tests of ensemble mean





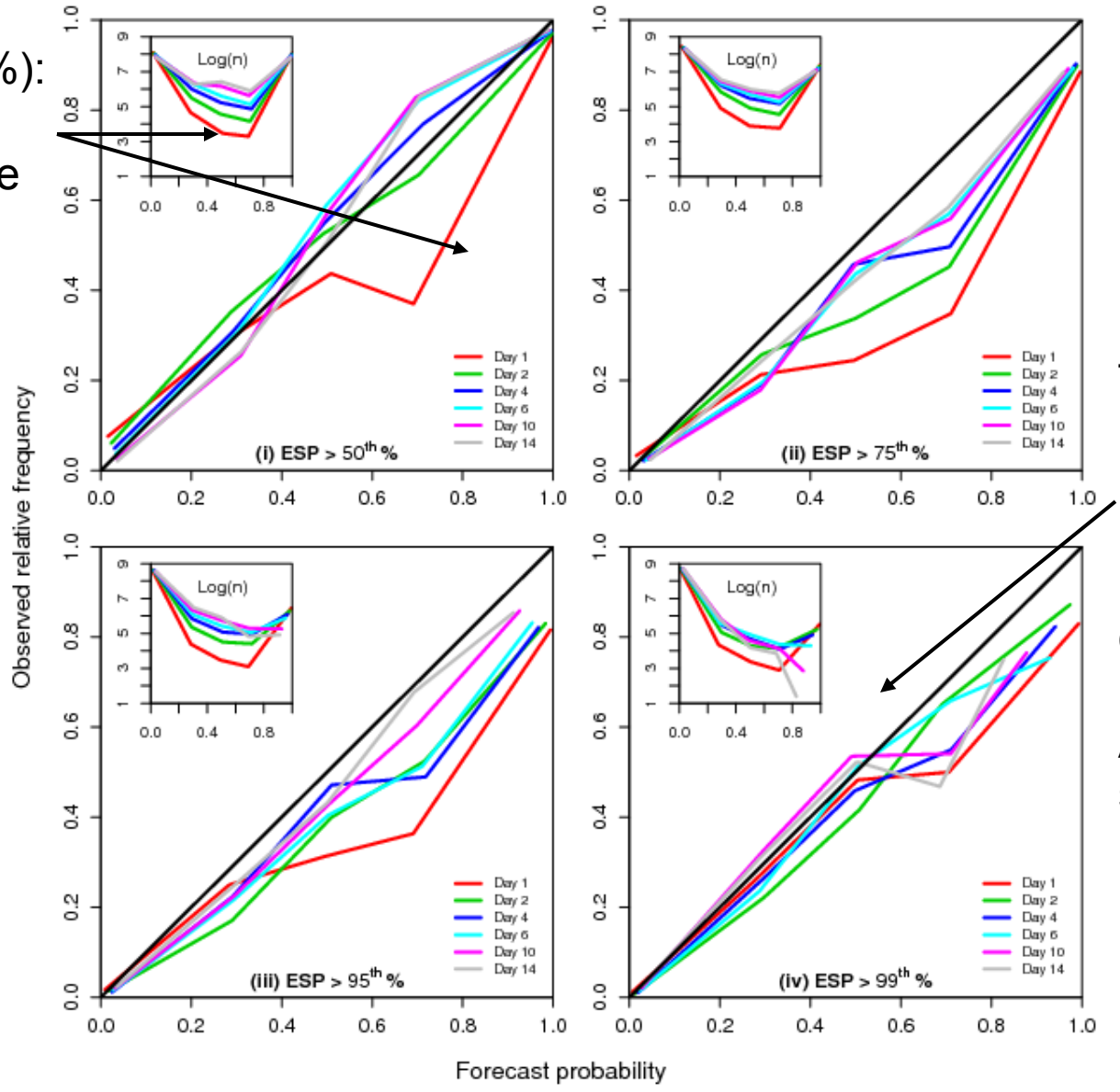


Lumped error in probability

Reliability

Day 1 (>50th%): sharp, but a little unreliable (contrast day 14).

No initial condition uncertainty (all forcing).



Day 14 (>99th%): forecasts remain reasonably reliable, but note 99% = only 210 CMS.

Also note sample size.

Next steps

To make EVS widely used (beyond NWS)

- Public download available (see next slide)
- Published in EM&S (others on apps.)

Ongoing research (two examples)

1) Verification of severe/rare events
   - Will benefit from new GEFS hindcasts
2) Detailed error source analysis
   - Hydrograph timing vs. magnitude errors (e.g. Cross-Wavelet Transform)

Hydrologic Ensemble Prediction
- About Us
- Projects
- Research and Development
- Experimental Applications and Products
- Publications and Presentations
- Events and Annoucements

Contact Us

The Ensemble Verification System (EVS)

[Revised: November 12, 2009]

Overview of the EVS

The EVS is designed for verifying ensemble forecasts of hydrologic and hydrometeorological variables, such as temperature, precipitation, streamflow, and river stage, issued at discrete forecast locations (points or areas). It is an experimental prototype developed by the Hydrological Ensemble Prediction group of the Office of Hydrologic Development.

The EVS is a Java application and will run on any operating system with a suitable Java Virtual Machine (it is currently tested on Linux and Windows). It is intended to be flexible, modular, and open to accommodate enhancements and additions by its developers and users. As such, we welcome your participation in the continuing development of the EVS toward a versatile and standardized tool for ensemble verification. The software is currently provided without technical support outside of the NWS. However, we welcome collaborations, suggestions and bug reports, for which a reporting template is provided with the distribution.

Reference Papers

- Brown, J.D., Demargne, J., Liu, Y. and Seo, D.J., 2009: The Ensemble Verification System (EVS): a software tool for verifying ensemble forecasts of hydrometeorological and hydrologic variables at discrete locations. Submitted to *Environmental Modelling and Software*.
- Demargne, J., Brown, J.D., Liu, Y., Seo, D.J., Wu, L., Toth, Z., and Zhu, Y., 2009: Diagnostic verification of hydrometeorological and hydrologic ensembles. Submitted to *Atmospheric Science Letters*.

Release History

- EVS Version 1.0: released in May 2008.
- EVS Version 2.0: released in October 2009.

Downloads

Version	Full Download	User's Manual	Executable	Release Notes	Test Data	Source Code	Developer Docs.	Bug Template
2.0 (10/16/09)	(25.4 MB)	(2.4 MB)	(13.3 MB)	(54 KB)	(1.7 MB)	(7.1 MB)	(1.5 MB)	(2 KB)

Disclaimer

This software and related documentation was developed by the National Weather Service (NWS). Pursuant to title 17, Section 105 of the United States Code this software is not subject to copyright protection and therefore may be used, copied, modified, and distributed without fee or cost. Parties who develop software incorporating predominantly NWS developed software must include notice as required by title 17, section 403 of the United States Code. NWS provides no warranty, expressed or implied, as to the correctness of the furnished software or the suitability for any purpose. NWS assumes no responsibility, whatsoever, for its use by other parties, about its quality, reliability, or any other characteristic. The NWS may change this software to meet its mission needs or discontinue its use without prior notice. The NWS cannot assist non-NWS users and is not obligated to fix reported problems; however, the NWS will make an attempt to fix reported problems where possible.

www.nws.noaa.gov/oh/evs.html

www.weather.gov/oh/XEFS/

Relevant published material.

**Full download;
user's manual (100 pp.);
source code; test data;
developer documentation
etc.**

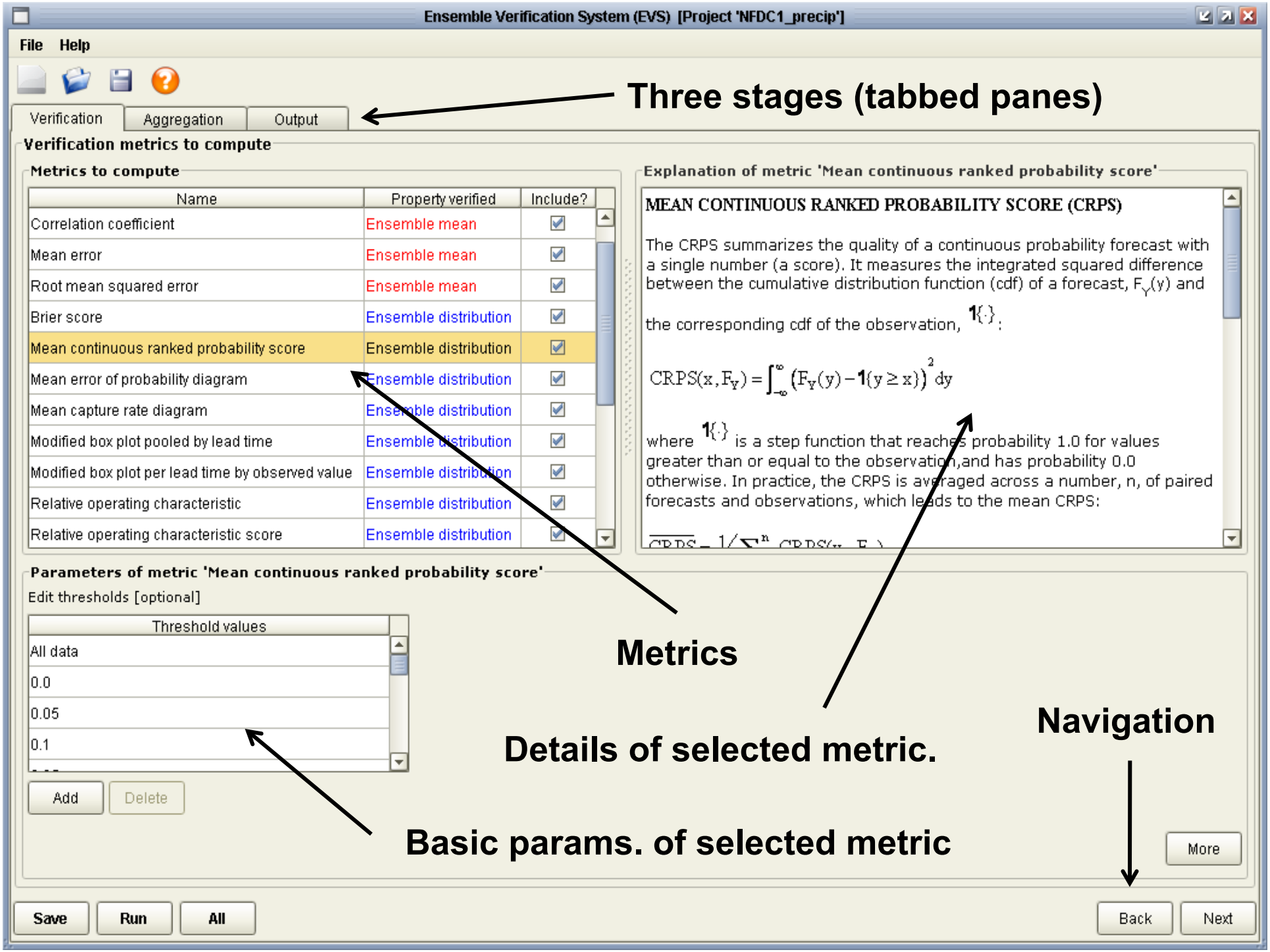
Follow-up literature

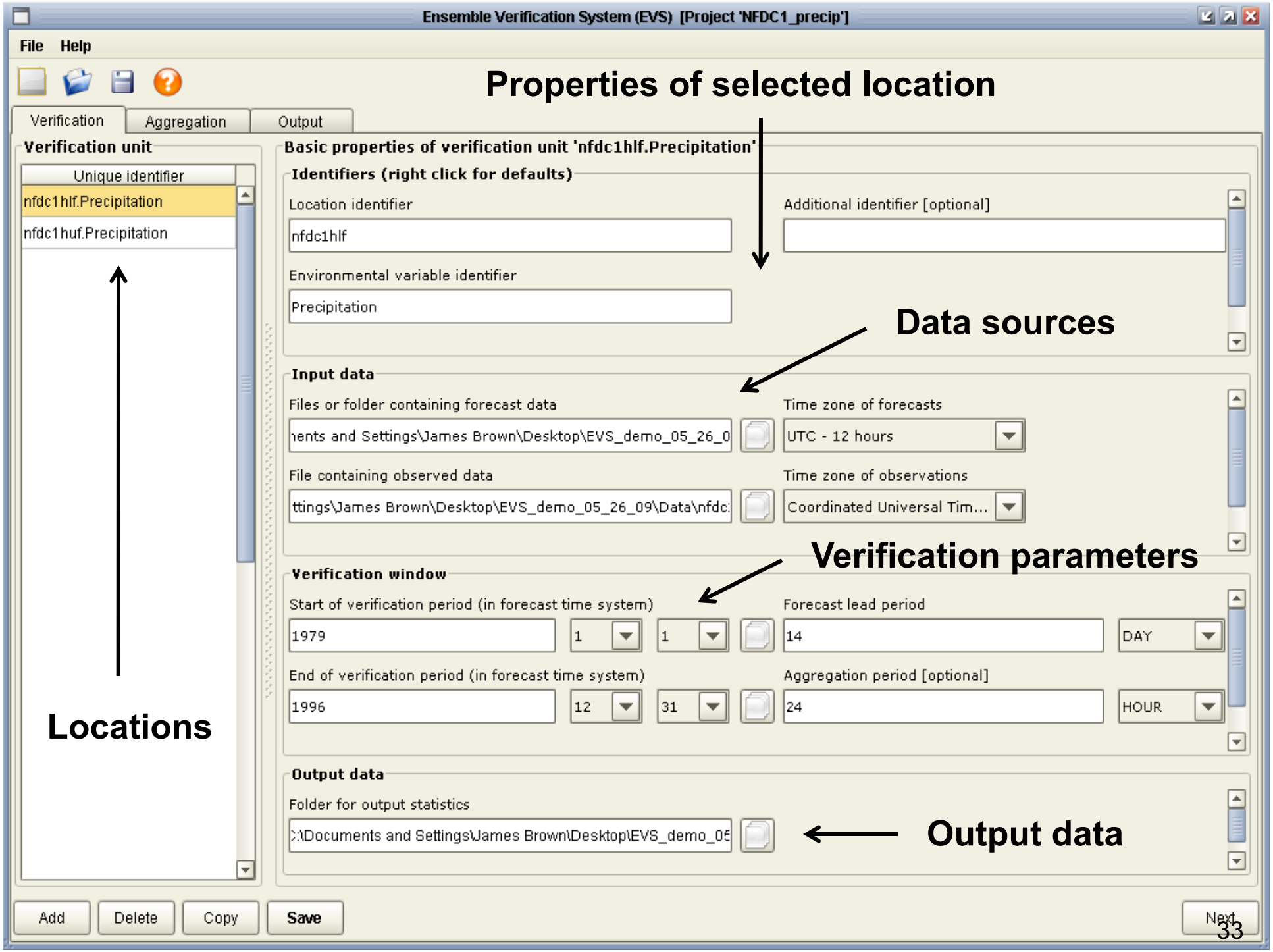
- Bradley, A. A., Schwartz, S. S. and Hashino, T., 2004: Distributions-Oriented Verification of Ensemble Streamflow Predictions. *Journal of Hydrometeorology*, **5(3)**, 532-545.
- Brown, J.D., Demargne, J., Liu, Y. and Seo, D-J (submitted) The Ensemble Verification System (EVS): a software tool for verifying ensemble forecasts of hydrometeorological and hydrologic variables at discrete locations. Submitted to *Environmental Modelling and Software.* 52pp.
- Gneiting, T., F. Balabdaoui, and Raftery, A. E., 2007: Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **69**(2), 243 – 268.
- Hsu, W.-R. and Murphy, A.H., 1986: The attributes diagram: A geometrical framework for assessing the quality of probability forecasts. *International Journal of Forecasting*, **2**, 285-293.
- Jolliffe, I.T. and Stephenson, D.B. (eds), 2003: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. Chichester: John Wiley and Sons, 240pp.
- Mason, S.J. and Graham N.E., 2002: Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation, *Quarterly Journal of the Royal Meteorological Society*, **30**, 291-303.
- Murphy, A. H. and Winkler, R.L., 1987: A general framework for forecast verification. *Monthly Weather Review*, **115**, 1330-1338.
- Wilks, D.S., 2006: *Statistical Methods in the Atmospheric Sciences, 2nd ed.* Academic Press, 627pp.

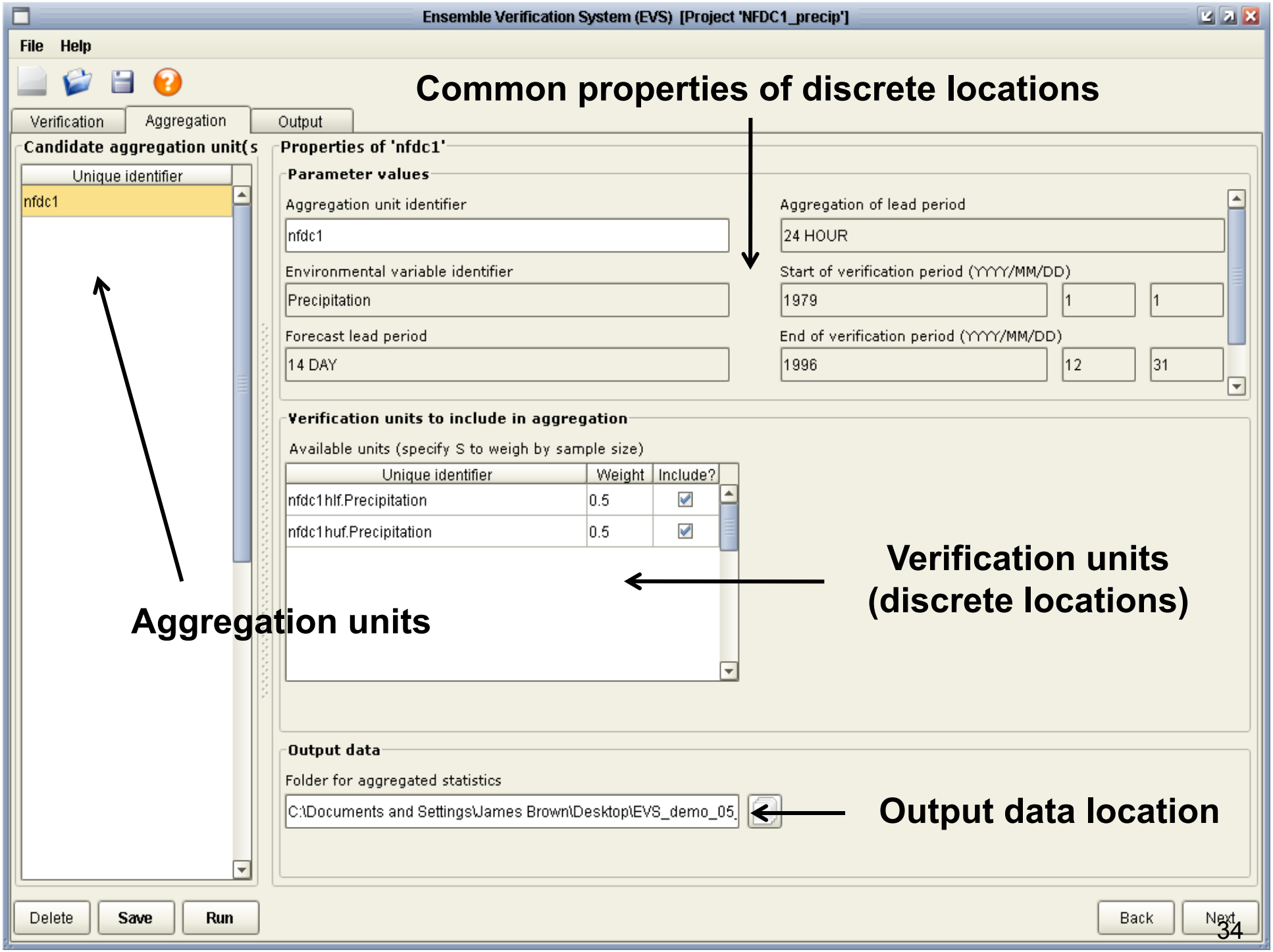
Additional slides

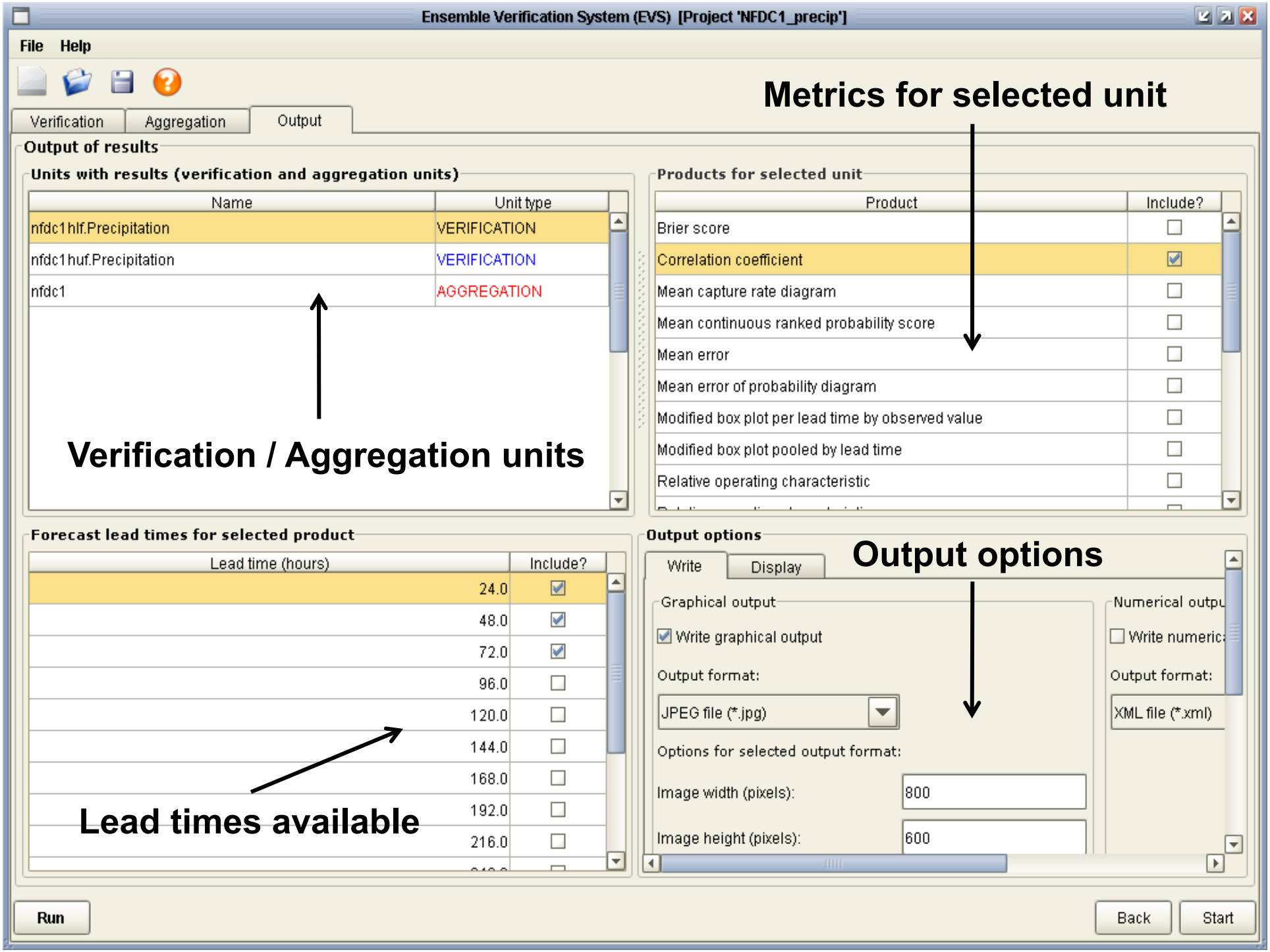
Verification metrics

Metric name	Quality tested	Discrete events?	Detail
Mean error	Ensemble mean	No	Lowest
RMSE	Ensemble mean	No	Lowest
Correlation coefficient	Ensemble mean	No	Lowest
Brier Score	Lumped error score	Yes	Low
Brier Skill Score	Lumped error score vs. reference	Yes	Low
Mean CRPS	Lumped error score	No	Low
Mean CRPS reliability	Lumped reliability score	No	Low
Mean CRPS resolution	Lumped resolution score	No	Low
CRPSS	Lumped error score vs. reference	No	Low
ROC score	Lumped discrimination score	Yes	Low
Mean error in prob.	Reliability (unconditional bias)	No	Low
Spread-bias diagram	Reliability (conditional bias)	No	High
Reliability diagram	Reliability (conditional bias)	Yes	High
ROC diagram	Discrimination	Yes	High
Modified box plots	Error visualization	No	Highest


Ensemble Verification System (EVS) [Project 'NFDC1_precip']
File Help
Three stages (tabbed panes)
Verification Aggregation Output
Verification metrics to compute
Metrics to compute
Name | Property verified | Include?
Correlation coefficient | Ensemble mean
Mean error | Ensemble mean
Root mean squared error | Ensemble mean
Brier score | Ensemble distribution
Mean continuous ranked probability score | Ensemble distribution
Mean error of probability diagram | Ensemble distribution
Mean capture rate diagram | Ensemble distribution
Modified box plot pooled by lead time | Ensemble distribution
Modified box plot per lead time by observed value | Ensemble distribution
Relative operating characteristic | Ensemble distribution
Relative operating characteristic score | Ensemble distribution
Explanation of metric 'Mean continuous ranked probability score'
MEAN CONTINUOUS RANKED PROBABILITY SCORE (CRPS)
The CRPS summarizes the quality of a continuous probability forecast with a single number (a score). It measures the integrated squared difference between the cumulative distribution function (cdf) of a forecast, F_Y(y) and the corresponding cdf of the observation, 1{.}:
CRPS(x, F_Y) = ∫_{-∞}^{∞} (F_Y(y) − 1(y ≥ x))^2 dy
where 1{.} is a step function that reaches probability 1.0 for values greater than or equal to the observation,and has probability 0.0 otherwise. In practice, the CRPS is averaged across a number, n, of paired forecasts and observations, which leads to the mean CRPS:
Parameters of metric 'Mean continuous ranked probability score'
Edit thresholds [optional]
Threshold values
All data
0.0
0.05
0.1
Add Delete
Metrics
Details of selected metric.
Navigation
Basic params. of selected metric
More
Save Run All
Back Next


Ensemble Verification System (EVS) [Project 'NFDC1_precip']
File Help
Properties of selected location
Verification
Aggregation
Output
Verification unit
Unique identifier
nfdc1hlf.Precipitation
nfdc1huf.Precipitation
Basic properties of verification unit 'nfdc1hlf.Precipitation'
Identifiers (right click for defaults)
Location identifier
nfdc1hlf
Additional identifier [optional]
Environmental variable identifier
Precipitation
Data sources
Input data
Files or folder containing forecast data
ients and Settings\James Brown\Desktop\EVS_demo_05_26_0
Time zone of forecasts
UTC - 12 hours
File containing observed data
ttings\James Brown\Desktop\EVS_demo_05_26_09\Data\nfdc
Time zone of observations
Coordinated Universal Tim...
Verification parameters
Verification window
Start of verification period (in forecast time system)
1979
1
1
Forecast lead period
14
DAY
End of verification period (in forecast time system)
1996
12
31
Aggregation period [optional]
24
HOUR
Output data
Folder for output statistics
:\Documents and Settings\James Brown\Desktop\EVS_demo_05
Output data
Locations
Add
Delete
Copy
Save
Next


Ensemble Verification System (EVS) [Project 'NFDC1_precip']
File Help
Common properties of discrete locations
Verification
Aggregation
Output
Candidate aggregation unit(s
Unique identifier
nfdc1
Properties of 'nfdc1'
Parameter values
Aggregation unit identifier
nfdc1
Environmental variable identifier
Precipitation
Forecast lead period
14 DAY
Aggregation of lead period
24 HOUR
Start of verification period (YYYY/MM/DD)
1979
1
1
End of verification period (YYYY/MM/DD)
1996
12
31
Verification units to include in aggregation
Available units (specify S to weigh by sample size)
Unique identifier
Weight
Include?
nfdc1hlf.Precipitation
0.5
nfdc1huf.Precipitation
0.5
Aggregation units
Verification units (discrete locations)
Output data
Folder for aggregated statistics
C:\Documents and Settings\James Brown\Desktop\EVS_demo_05
Output data location
Delete
Save
Run
Back
Next


Ensemble Verification System (EVS) [Project 'NFDC1_precip']
File Help
Verification Aggregation Output
Output of results
Units with results (verification and aggregation units)
Name
Unit type
nfdc1 hlf.Precipitation
VERIFICATION
nfdc1 huf.Precipitation
VERIFICATION
nfdc1
AGGREGATION
Verification / Aggregation units
Metrics for selected unit
Products for selected unit
Product
Include?
Brier score
Correlation coefficient
Mean capture rate diagram
Mean continuous ranked probability score
Mean error
Mean error of probability diagram
Modified box plot per lead time by observed value
Modified box plot pooled by lead time
Relative operating characteristic
Forecast lead times for selected product
Lead time (hours)
Include?
24.0
48.0
72.0
96.0
120.0
144.0
168.0
192.0
216.0
Lead times available
Output options
Write
Display
Output options
Graphical output
Write graphical output
Output format:
JPEG file (*.jpg)
Options for selected output format:
Image width (pixels):
800
Image height (pixels):
600
Numerical outpu
Write numeric
Output format:
XML file (*.xml)
Run
Back
Start